Deep learning

Dr. Aissa Boulmerka
a.boulmerka@centre-univ-mila.dz

2023-2024

CHAPTER 9
SEQUENCE MODELS & ATTENTION MECHANISM

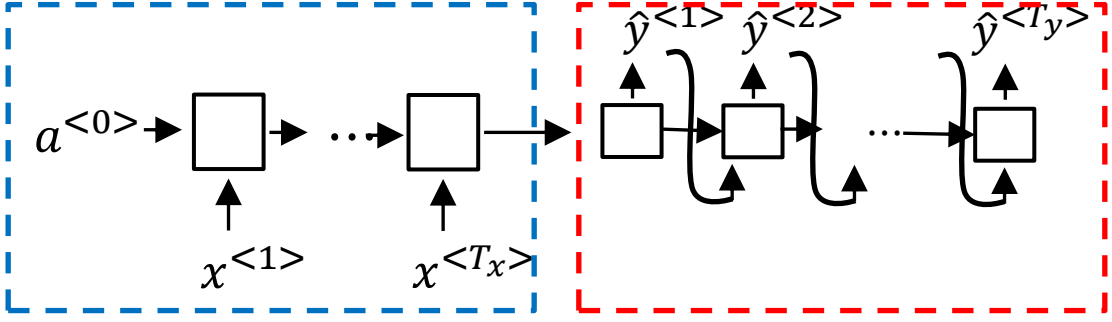
Sequence to sequence model

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ $x^{<4>}$ $x^{<5>}$

Jane visite l'Afrique en septembre

→ Jane is visiting Africa in September.

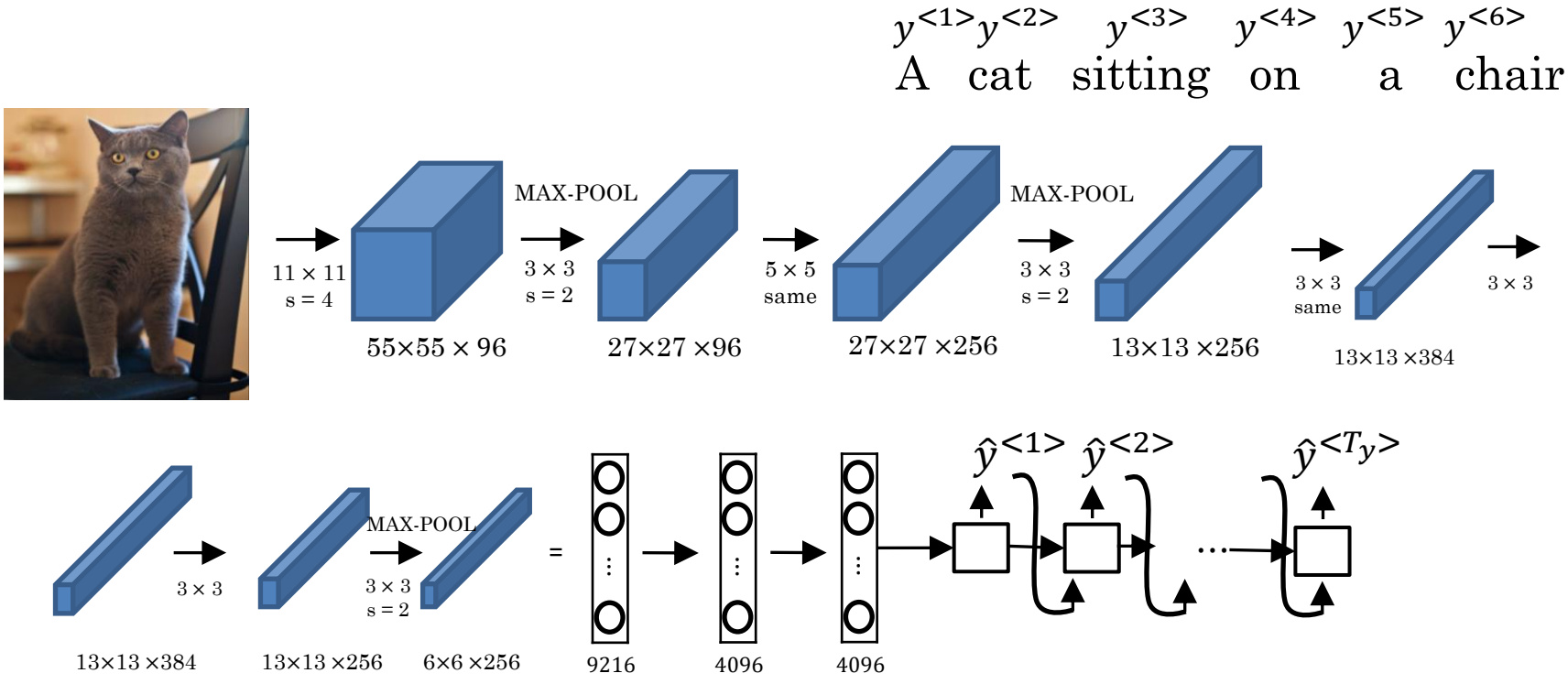
$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ $y^{<4>}$ $y^{<5>}$ $y^{<6>}$



[Sutskever et al., 2014. Sequence to sequence learning with neural networks]

[Cho et al., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation]

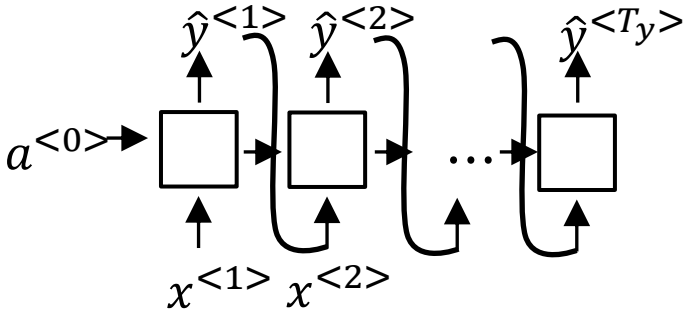
Image captioning



[Mao et. al., 2014. Deep captioning with multimodal recurrent neural networks]
[Vinyals et. al., 2014. Show and tell: Neural image caption generator]
[Karpathy and Li, 2015. Deep visual-semantic alignments for generating image descriptions]

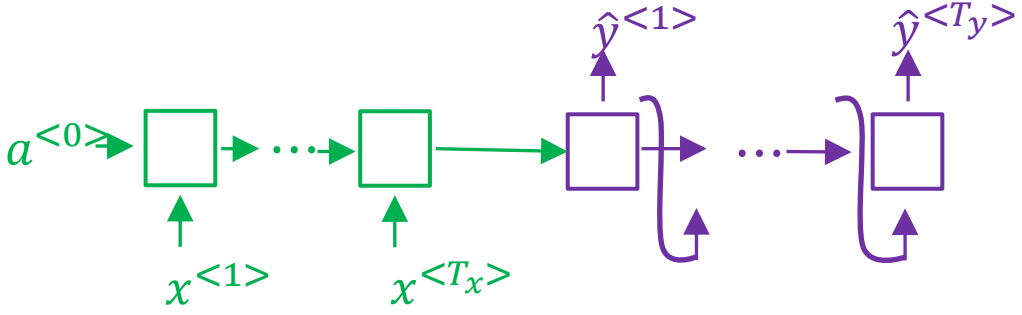
Machine translation as building a conditional language model

Language model:



$P(y^{<1>}, \cdots, y^{<T_y>})$

Machine translation:



Conditional language model $P(y^{<1>}, \cdots, y^{<T_y>} | x^{<1>}, \cdots, x^{<T_x>})$

PICKING THE MOST LIKELY SENTENCE

Finding the most likely translation

English **French**

Jane visite l'Afrique en septembre. $P(y^{<1>}, \ldots, y^{<T_y>} | x)$

→ Jane is visiting Africa in September.

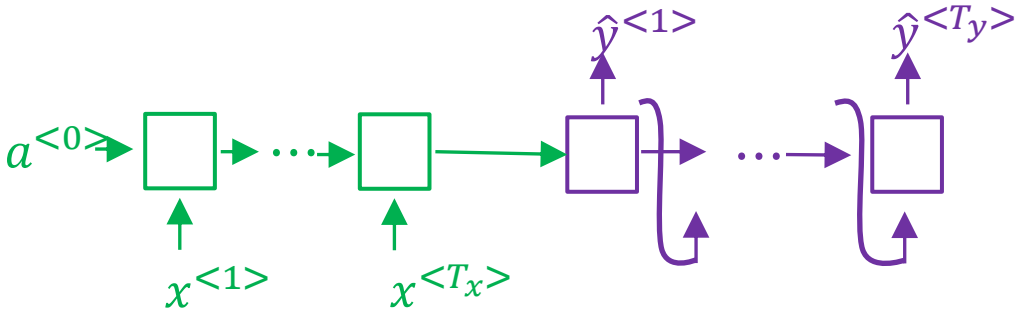
→ Jane is going to be visiting Africa in September.

→ In September, Jane will visit Africa.

→ Her African friend welcomed Jane in September.

$$\underset{y^{<1>},\ldots,y^{<T_y>}}{\arg\max} \; P(y^{<1>}, \ldots, y^{<T_y>} | x)$$

Why not a greedy search?



$$\arg\max_{y^{<1>},\ldots,y^{<T_y>}} P(y^{<1>},\ldots,y^{<T_y>}|\, x)$$

→ Jane is visiting Africa in September.

→ Jane is going to be visiting Africa in September.

$P(\text{Jane is going}|\, x) > P(\text{Jane is visiting}|\, x)$!!!

BEAM SEARCH

Beam search algorithm

Step 1

B=3 (beam width)

10000

a

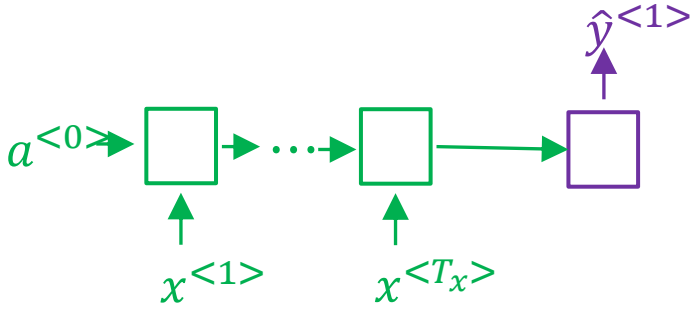
in

jane

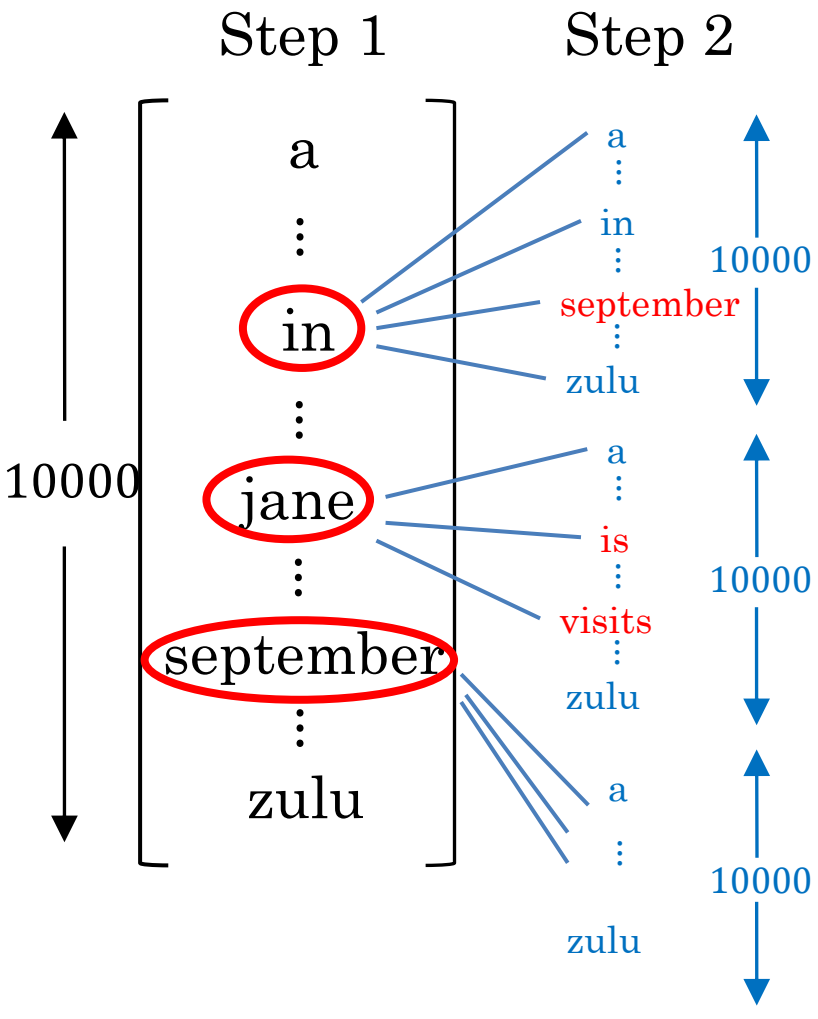
september

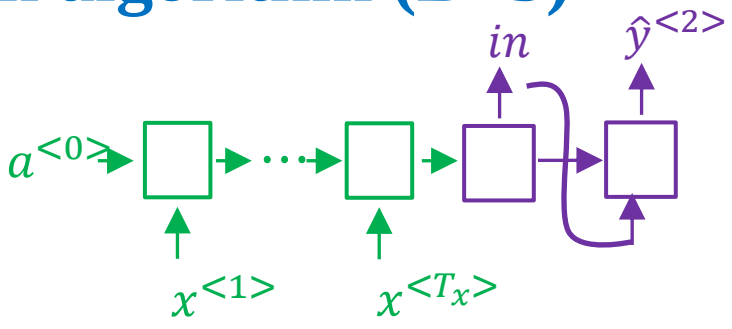
zulu

$P(y^{<1>} \mid x)$

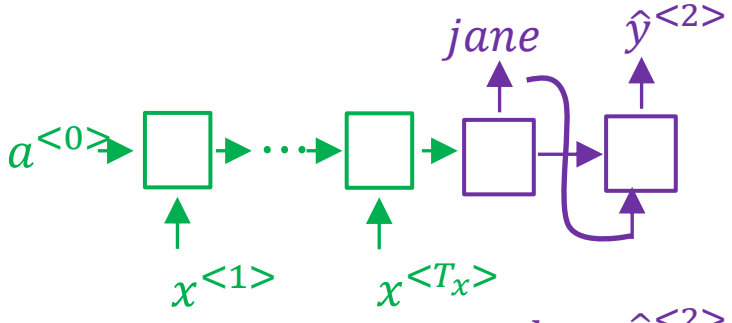


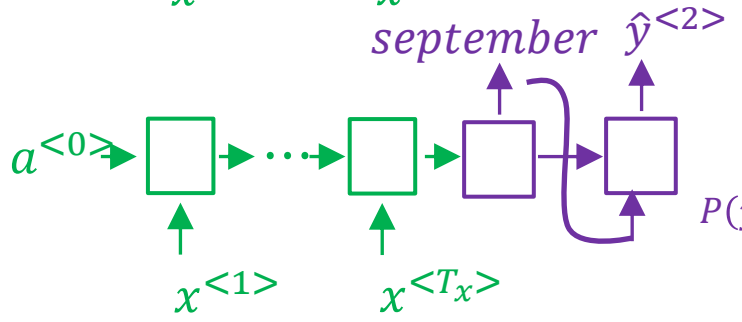
Beam search algorithm (B=3)



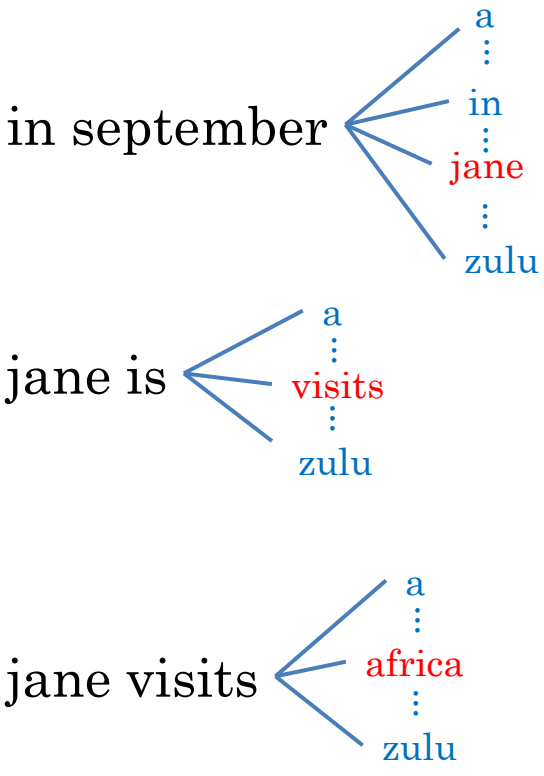


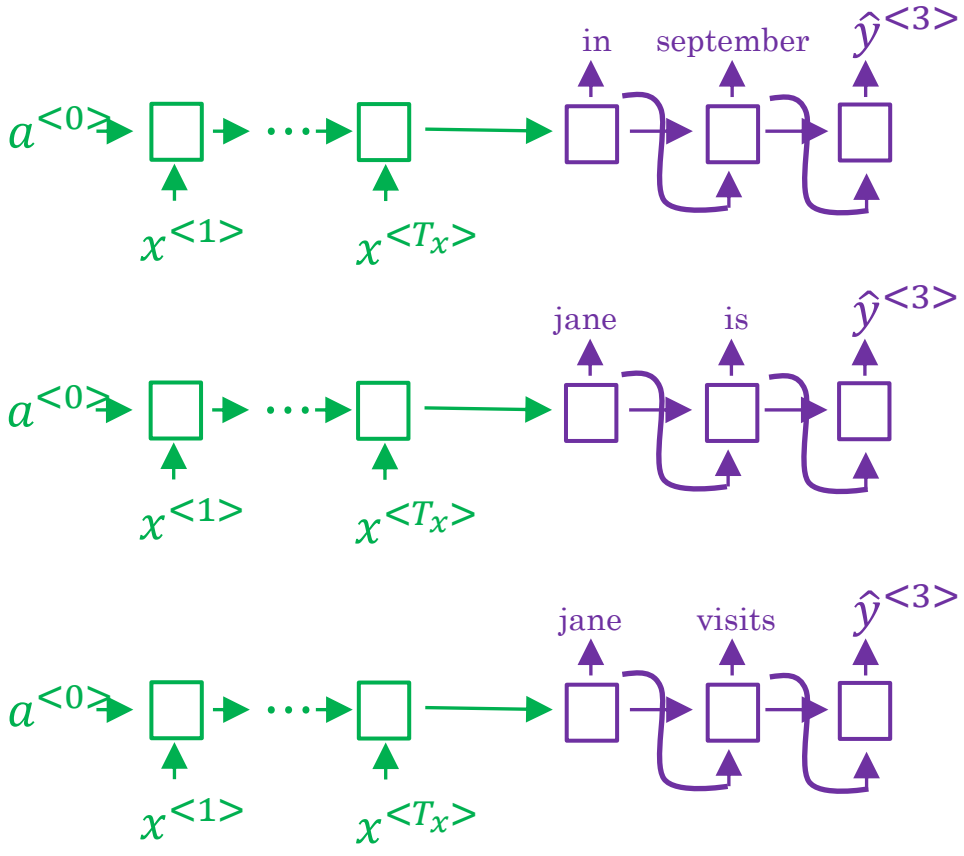
$$P(y^{<1>}, y^{<2>}|\, x) = P(y^{<1>}\,|\, x)P(y^{<2>}|x, y^{<1>})$$





Beam search ($B = 3$)





$P(y^{<1>}, y^{<2>} | x)$

jane visits africa in september. <EOS>

REFINEMENTS TO BEAM SEARCH

Length normalization

$$\arg\max_{y} \prod_{t=1}^{T_y} P(y^{<t>} | x, y^{<1>}, \dots, y^{<t-1>})$$

$$\arg\max_{y} \sum_{t=1}^{T_y} \log P(y^{<t>} | x, y^{<1>}, \dots, y^{<t-1>})$$

$$\frac{1}{{T_y}^{\alpha}} \sum_{t=1}^{T_y} \log P(y^{<t>} | x, y^{<1>}, \dots, y^{<t-1>})$$

Beam search discussion

Beam width B?

Large B: better result , slower

Small B: worse result, faster

Unlike exact search algorithms like BFS (Breadth First Search) or DFS (Depth First Search), Beam Search runs faster but is not guaranteed to find exact maximum for $\arg\max_y P(y|x)$.

ERROR ANALYSIS IN BEAM SEARCH

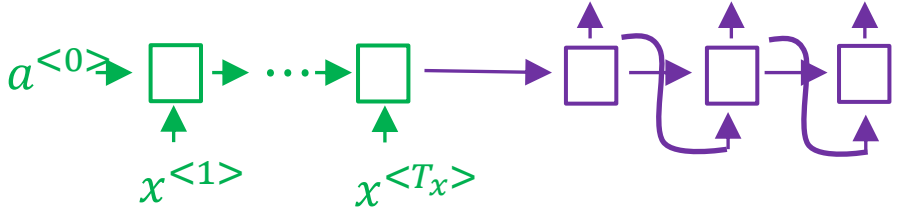
Example

Jane visite l'Afrique en septembre.

Human: Jane visits Africa in September.

Algorithm: Jane visited Africa last September.

$$P(y^*|x) > P(\hat{y}|x)?$$



Error analysis on beam search

Human: Jane visits Africa in September. (y^*) $P(y^*|x)$

Algorithm: Jane visited Africa last September. ($\hat{y}$) $P(\hat{y}|x)$

Case 1: $P(y^*|x) > P(\hat{y}|x)$

Beam search chose $\hat{y}$. But y^* attains higher $P(y|x)$.

Conclusion: Beam search is at fault.

Case 2: $P(y^*|x) \leq P(\hat{y}|x)$

y^* is a better translation than $\hat{y}$. But RNN predicted $P(y^*|x) < P(\hat{y}|x)$.

Conclusion: RNN model is at fault.

Error analysis process

Human	Algorithm	$P(y^*\|x)$	$P(\hat{y}\|x)$	At fault?
Jane visits Africa in September.	Jane visited Africa last September.			

Figures out what faction of errors are “due to” beam search vs. RNN model

BLEU SCORE

Evaluating machine translation

French: Le chat est sur le tapis.

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

MT output: the the the the the the the.

Precision: Modified precision:

[Papineni et. al., 2002. Bleu: A method for automatic evaluation of machine translation]

Bleu score on bigrams

Example: Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

MT output: The cat the cat on the mat.

the cat

cat the

cat on

on the

the mat

[Papineni et. al., 2002. Bleu: A method for automatic evaluation of machine translation]

Bleu score on unigrams

Example: Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

MT output: The cat the cat on the mat.

$$p_1 = \frac{\sum_{unigram \in \hat{y}} count_{clip}\,(unigram)}{\sum_{unigram \in \hat{y}} count\,(unigram)} \qquad p_n = \frac{\sum_{ngram \in \hat{y}} count_{clip}\,(ngram)}{\sum_{ngram \in \hat{y}} count\,(ngram)}$$

[Papineni et. al., 2002. Bleu: A method for automatic evaluation of machine translation]

Bleu details

p_n = Bleu score on n-grams only

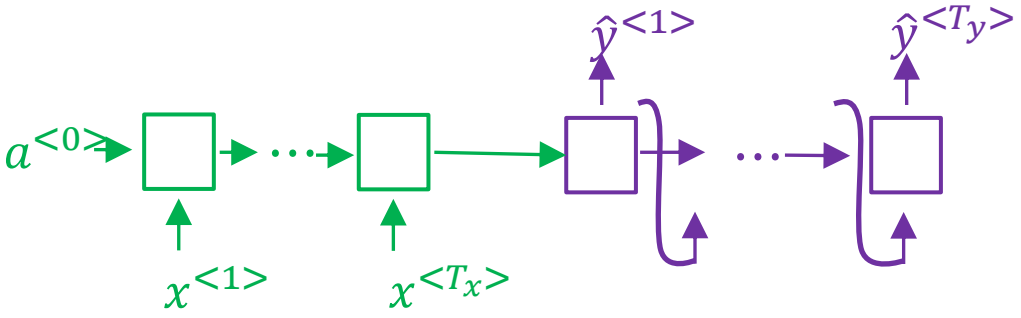
Combined Bleu score:

$$\text{BP} = \begin{cases} 1 & \text{if MT\_output\_length} > \text{reference\_output\_length} \\ \exp(1 - \text{MT\_output\_length}/\text{reference\_output\_length}) & \text{otherwise} \end{cases}$$

[Papineni et. al., 2002. Bleu: A method for automatic evaluation of machine translation]

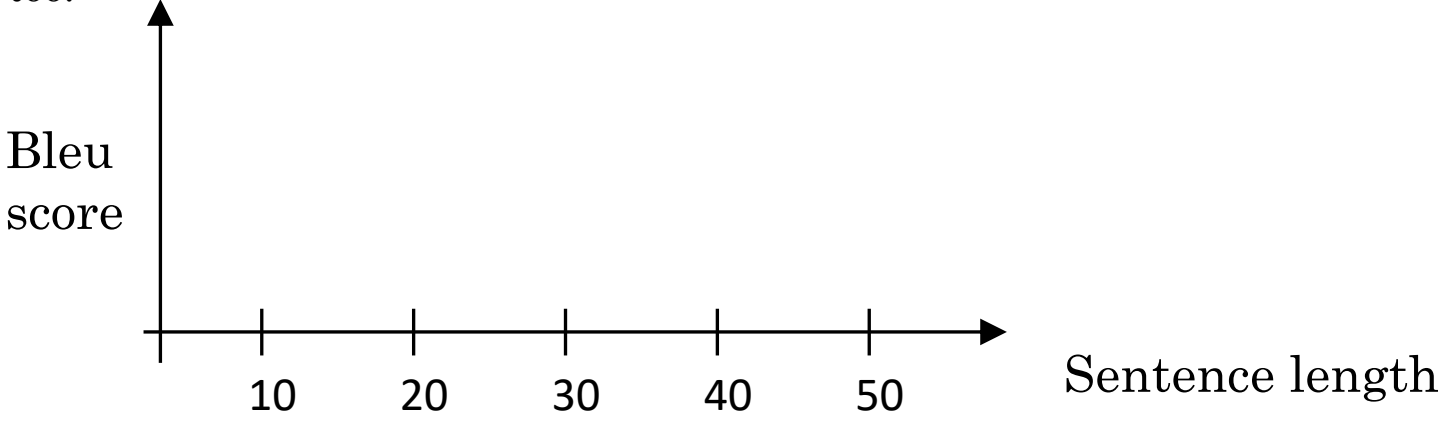
ATTENTION MODEL INTUITION

The problem of long sequences

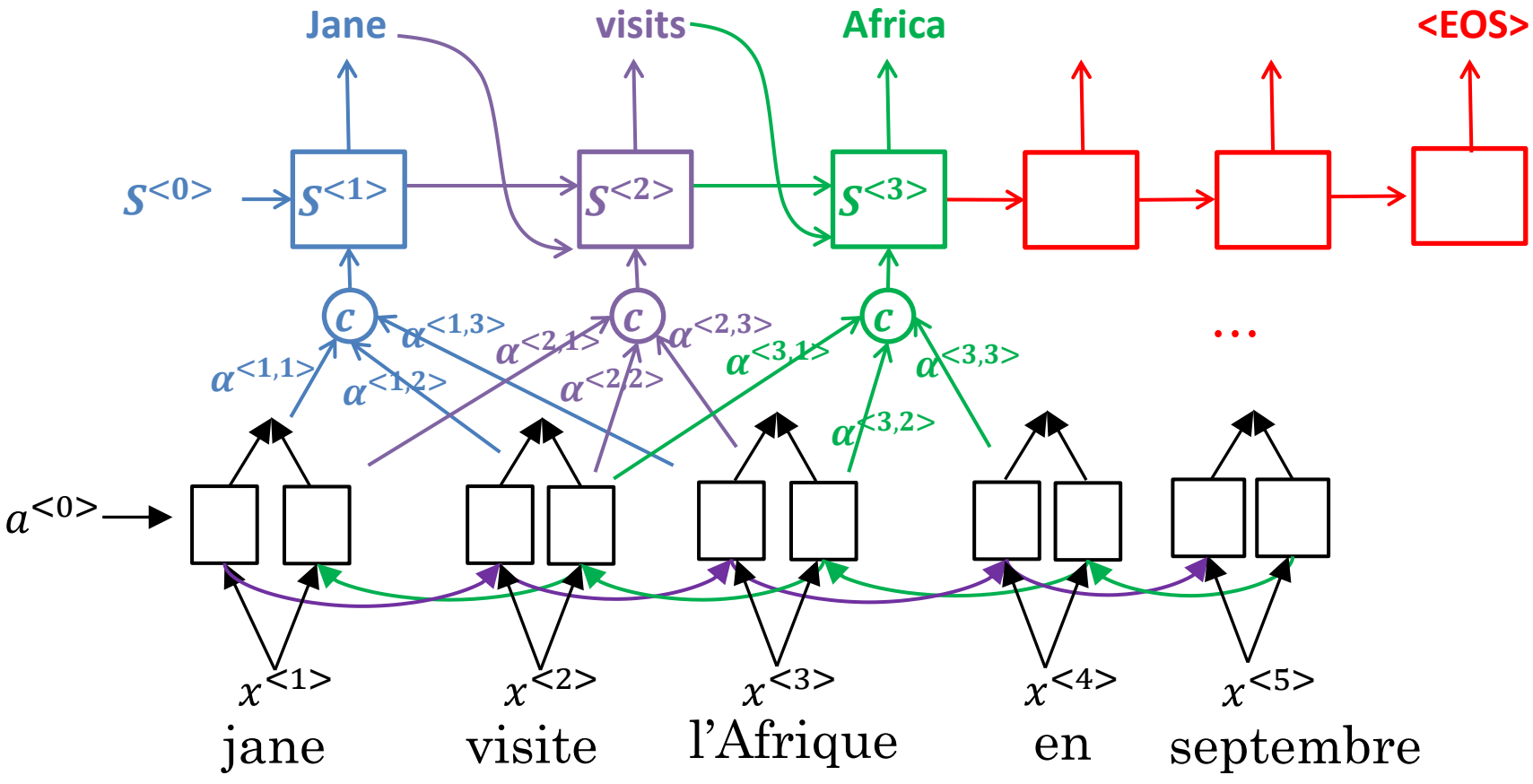


Jane s'est rendue en Afrique en septembre dernier, a apprécié la culture et a rencontré beaucoup de gens merveilleux; elle est revenue en parlant comment son voyage était merveilleux, et elle me tente d'y aller aussi.

Jane went to Africa last September, and enjoyed the culture and met many wonderful people; she came back raving about how wonderful her trip was, and is tempting me to go too.

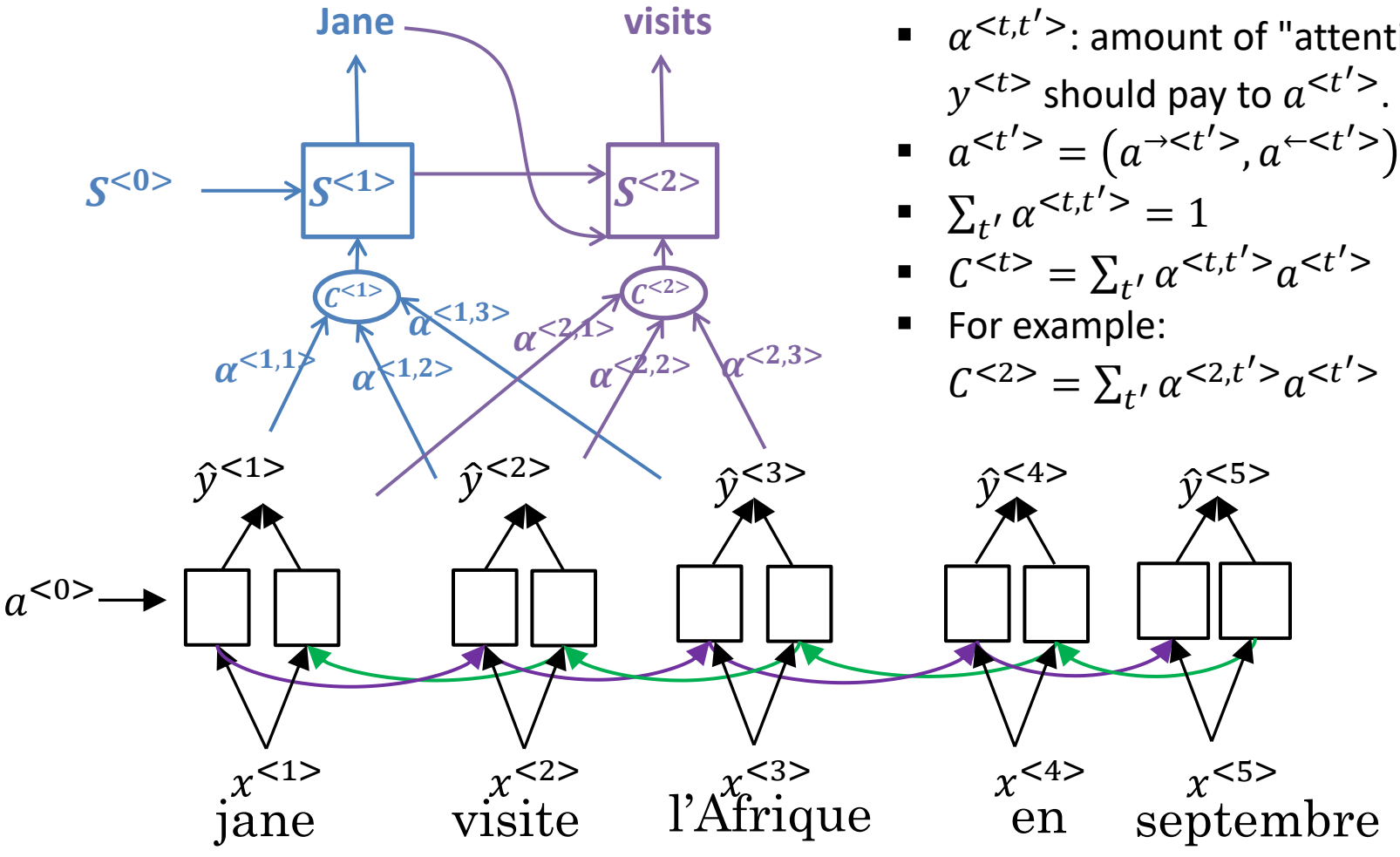


Attention model



[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]

Attention model intuition



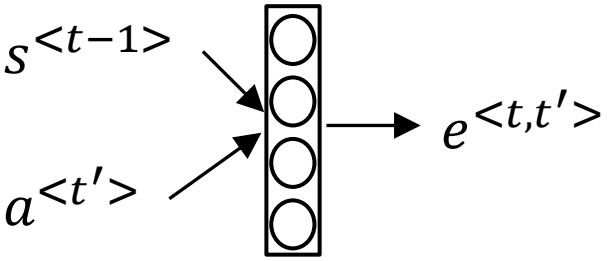
- $\alpha^{<t,t'>}$: amount of "attention" $y^{<t>}$ should pay to $a^{<t'>}$.
- $a^{<t'>} = \left(a^{\rightarrow<t'>}, a^{\leftarrow<t'>}\right)$
- $\sum_{t'} \alpha^{<t,t'>} = 1$
- $C^{<t>} = \sum_{t'} \alpha^{<t,t'>} a^{<t'>}$
- For example: $C^{<2>} = \sum_{t'} \alpha^{<2,t'>} a^{<t'>}$

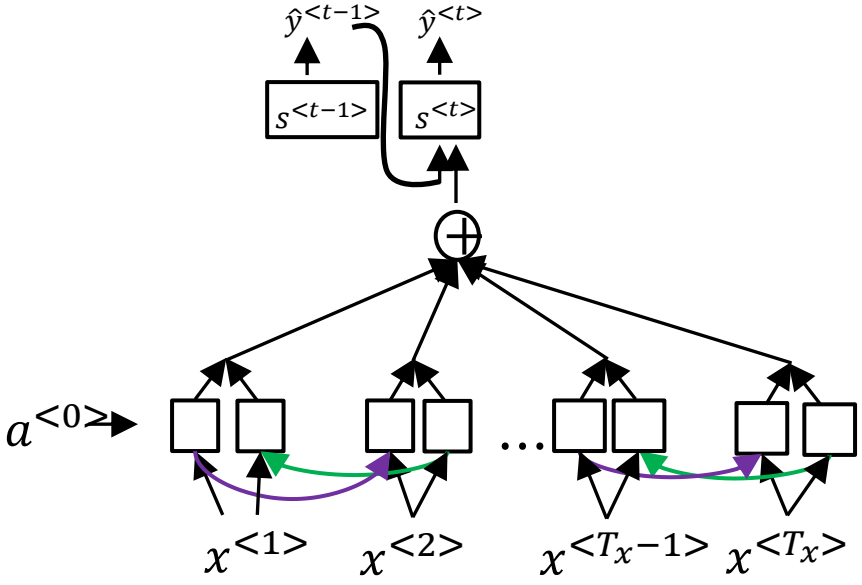
[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]

Computing attention $\alpha^{<t,t'>}$

$\alpha^{<t,t'>}$ = amount of attention $y^{<t>}$ should pay to $a^{<t'>}$

$$\alpha^{<t,t'>} = \frac{\exp(e^{<t,t'>})}{\sum_{t'=1}^{T_x} \exp(e^{<t,t'>})}$$





[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]

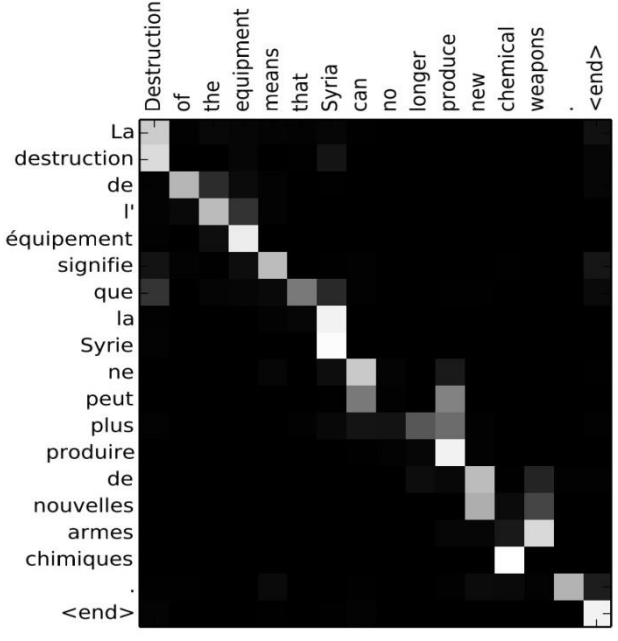
[Xu et. al., 2015. Show, attend and tell: Neural image caption generation with visual attention]

Attention examples

July 20th 1969 $\longrightarrow$ 1969 – 07 – 20

23 April, 1564 $\longrightarrow$ 1564 – 04 – 23

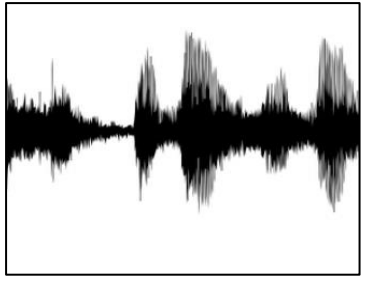
Visualization of $\alpha^{<t,t'>}$:



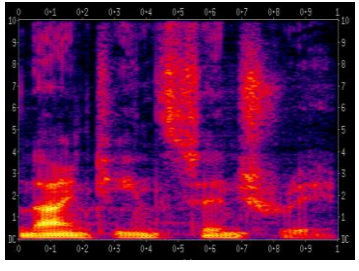
Speech recognition problem

x → y

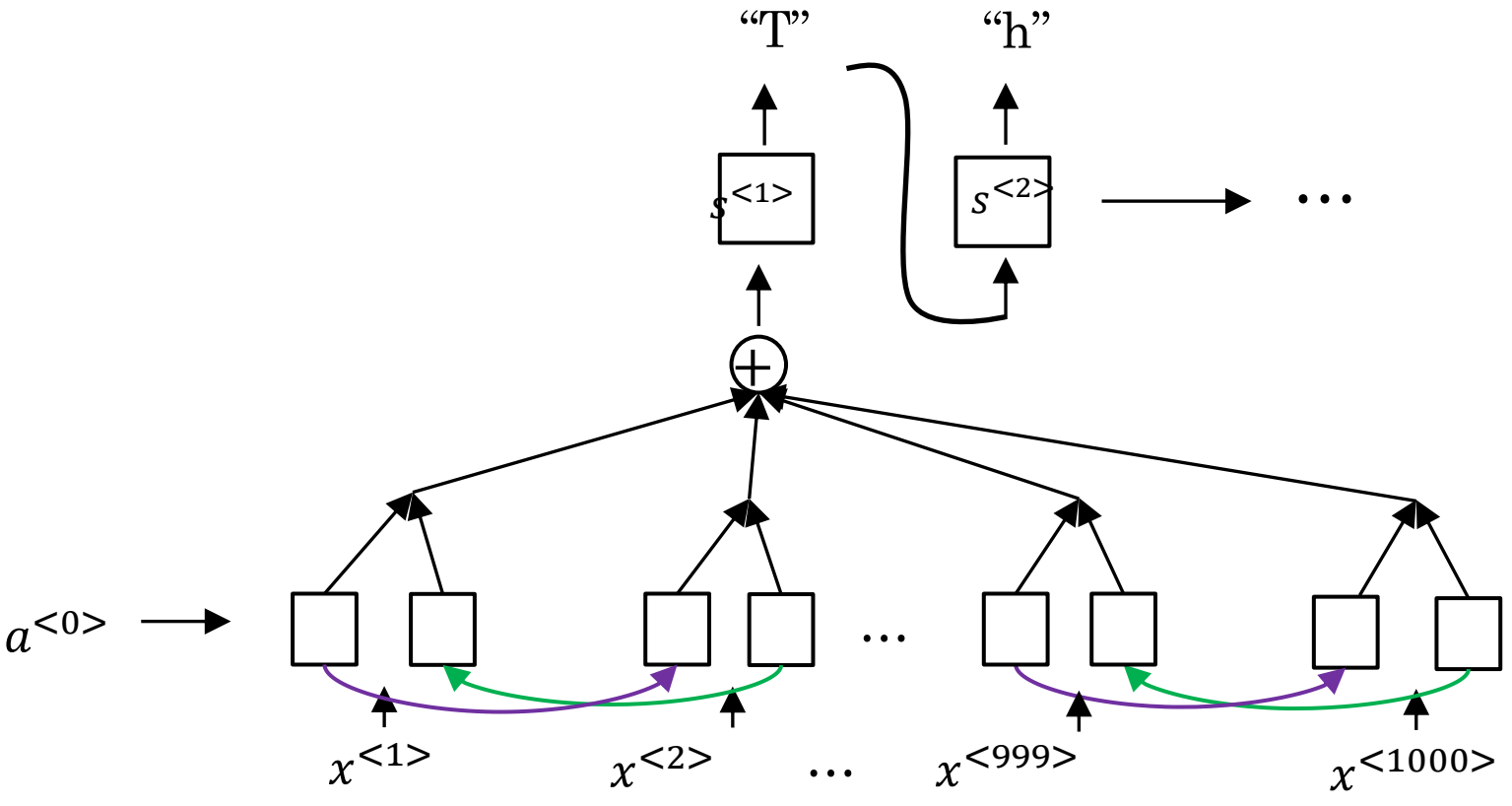
audio clip → transcript



→ “the quick brown fox”



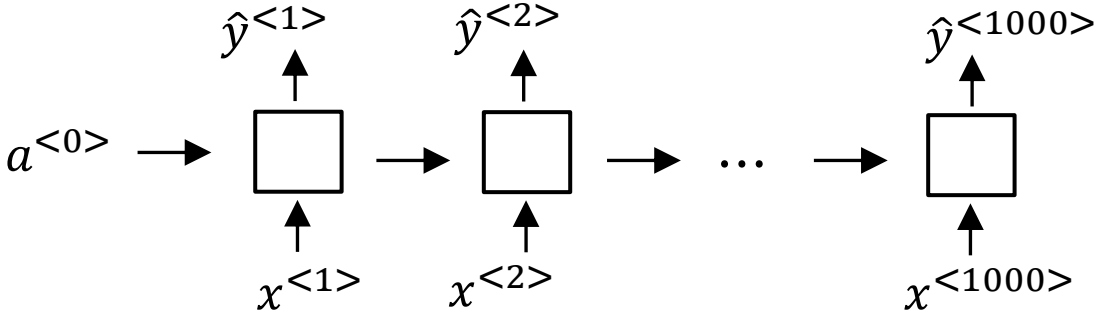
Attention model for speech recognition



CTC cost for speech recognition

(Connectionist temporal classification)

"the quick brown fox"



Basic rule: collapse repeated characters not separated by "blank"

[Graves et al., 2006. Connectionist Temporal Classification: Labeling unsegmented sequence data with recurrent neural networks]

What is trigger word detection?

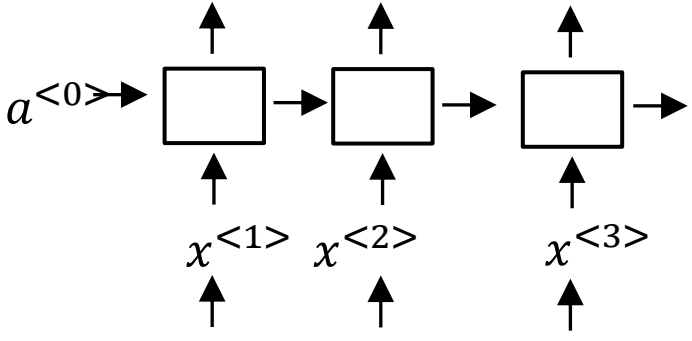
Amazon Echo
(Alexa)

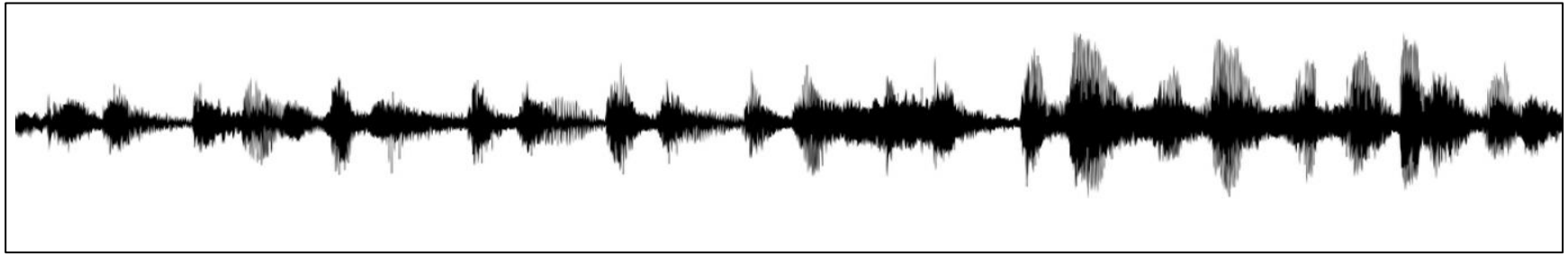
Baidu DuerOS
(xiaodunihao)

Apple Siri
(Hey Siri)

Google Home
(Okay Google)

Trigger word detection algorithm





References

- Andrew Ng. Deep learning. Coursera.
- Geoffrey Hinton. Neural Networks for Machine Learning.
- Kevin P. Murphy. Probabilistic Machine Learning An Introduction. MIT Press, 2022.
- MIT Deep Learning 6.S191 (http://introtodeeplearning.com/)